How do pessimistic agents save miners? A STIT based approach

Xin Sun[1], Zohreh Baniasadi[2], Shuwen Zhou[3]

[1,2]Faculty of Science, Technology and Communication, University of Luxembourg,
[3] School of computer Science Engineering, University of New South Wales,
[1]xin.sun@uni.lu, [2]zohreh.baniasadi.001@student.uni.lu, [3]Jacobc3@gmail.com

Abstract. This paper develops a new STIT based deontic logic, pessimistic utilitarian deontic logic, capable of analyzing the miners puzzle. The key idea of the semantics of this logic is: one set of possible worlds is better than another set of possible worlds iff the worst world in the first set is better than the worst world in the second. This semantics gives an idea to write predictions in the miners scenario meanwhile blocks the unwilling statement.

1 Introduction

This paper develops a new STIT based deontic logic, referring it as pessimistic utilitarian deontic logic, capable of analyzing a recently popular puzzle in deontic logic, the miners puzzle [9]. The miners puzzle goes like this:

> Ten miners are trapped either in shaft A or in shaft B, but we do not know which one. Water threatens to flood the shafts. We only have enough sandbags to block one shaft but not both. If one shaft is blocked, all of the water will go into the other shaft, killing every miner if they are inside. If we block neither shaft, both will be partially flooded, killing one miner.

Lacking any information about the miners' exact whereabouts, it seems acceptable to say that:

(1) We ought to block neither shaft.

However, we also accept that

(2) If the miners are in shaft A, we ought to block shaft A.

(3) If the miners are in shaft B, we ought to block shaft B.

But we also know that

(4) Either the miners are in shaft A or they are in shaft B.

And (2)-(4) seem to entail

(5) Either we ought to block shaft A or we ought to block shaft B.

Which contradicts (1).

It is stated by Willer [16], any adequate semantics of dyadic deontic modality must offer a solution to the miners puzzle. The existing STIT-based deontic logic [7, 11, 14] does not offer a satisfying analysis to this puzzle: although the deduction from (2)-(4) to (5) is blocked by the dyadic deontic operator defined in Sun [14], but both Horty [7] and Sun [14] are unable to predict (1). In this paper our motivation is to develop a new STIT-based deontic logic which is capable of blocking the deduction from (2)-(4) to (5) and it is able to predict (1)-(4).

In STIT-based deontic logic, agents make choices. The outcome of agents' choice is represented by a set of possible worlds. A preference relation over the set of all possible worlds is given as primitive. This preference relation is then lifted to the relation of preference over sets of possible worlds. A choice is better than another iff the resulting set of worlds of the first choice is better than the resulting set of worlds of the second. A proposition φ is obligatory (we ought to see to it that φ) iff it is ensured by every best choice, i.e., it is true in every world of every best choice.

Therefore the interpretation of deontic modality is based on best choices, which can only be defined on top of preference over sets of worlds, which is defined by lifting from the preference over worlds. There is no standard way of lifting preference. Lang and van der Torre [13] discuss the following three ways of lifting:

- **strong lifting** For two sets of worlds W_1 and W_2, W_1 is strongly better than W_2 iff $\forall w \in W_1, \forall v \in W_2$, w is better than v. That is, the worst world in W_1 is better than the best world in W_2.
- **optimistic lifting** W_1 is optimistically better than W_2 iff $\exists w \in W_1, \forall v \in W_2$, w is better than v. That is, the best world in W_1 is better than the best world in W_2.
- **pessimistic lifting** W_1 pessimistically better than W_2 iff $\forall w \in W_1, \exists v \in W_2$, w is better than v. That is, the worst world in W_1 is better than the worst world in W_2.

In Horty [7], Kooi and Tamminga [11] and Sun [14] the strong lifting is adopted. Applying the strong lifting to the miners scenario, all the three choices *block_neither*, *block*_A and *block*_B are best. "we ought to block neither" is then not true in the miners scenario in the logic Horty [7], Kooi and Tamminga [11] and Sun [14].

In this paper we use pessimistic lifting instead of strong lifting. There is a single best choice *block_neither* according to pessimistic lifting. Therefore "we ought to block neither" is true. It can be further proved that both (2) and (3) are true while the deduction from (2)-(4) to (5) is not valid. Therefore our logic offers a satisfying solution to the miners paradox.

The structure of this paper is as following: in Section 2 we review the existing solutions to the miners puzzle. Then in Section 3 we review the existing STIT-based deontic logic. In Section 4 we develop the pessimistic utilitarian deontic logic and offer a viable solution to the miners puzzle. Section 5 is our conclusion and future work.

2 Solutions of the miners paradox

Several authors have provided different solutions to solve the miners puzzle. Among them, we summarize the following approaches:

Kolodny and MacFarlane [9] give a detailed discussion of various escape routes. Then they conclude that the only possible solution to the puzzle is to invalidate the argument from (2) to (5). To do this, Kolodny and MacFarlane state we have three choices: rejecting modus ponens (MP), rejecting disjunction introduction ($\vee$I), rejecting disjunction elimination ($\vee$E). Among these three Kolodny and MacFarlane further demonstrate that the only wise choice is to reject MP.

Willer [16] developed a fourth option to invalidate the argument form (2) to (5): falsify the monotonicity. In his solution the modus ponens can be preserved (there are very good reasons to do so) and we are unable to derive the inconsistency.

Charlow [5] proposes a comprehensive solution which requires rethinking the relationship between relevant information (what we know) and practical rankings of possibilities and actions (what to do).

Cariani et al [3] argues that the traditional Kratzer's semantics [12] of deontic conditionals is not capable of solving the puzzle. They propose to extend the standard Kratzer's account by adding a parameter representing a "decision problem" to solve the puzzle. The "decision problem" shares some similarity to STIT operator with single agent.

Carr [4] argues that the proposal of Cariani et al is still problematic. To develop a satisfying semantics, Carr uses three parameters to define deontic modality: an informational parameter, a value parameter and a decision rule parameter. According to Carr's proposal, (1) to (3) are all correct predictions and no contradiction arise within her framework.

Gabbay et al [6] offers a solution to the miners puzzle using idea from intuitionistic logic. In their logic "or" is interpreted in an intuitionistic favour. Then the reasoning from statement (2) to (5) is blocked.

3 STIT-based deontic logic

In this section we review STIT-based deontic logic. Following Horty [7], we call such logic utilitarian deontic logic (UDL).

3.1 Language

The language of the UDL is built from a finite set $Agent$ of agents and a countable set P of propositional letters. We will use p, q as variables for atomic propositions in P, and use G, with $G \subseteq Agent$, as a group of agents. The utilitarian deontic language L is given by the following Backus-Naur Form:

$$\varphi \ ::= \ p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \bigcirc_G \varphi \mid \bigcirc_G(\varphi/\varphi)$$

Intuitively, $\bigcirc_G \varphi$ is read as "G ought to see to it that φ". $\bigcirc_G(\varphi/\psi)$ is read as "G ought to see to it that φ under the condition ψ".

Our language is simpler than the language of Horty [7] in the sense that we omit the possibility operator $\Diamond$ and we represent the ought operator in a more compact way than Horty [7].

3.2 Semantics

The semantics of utilitarian deontic logic is based on the utilitarian frames, which is a simplification of STIT frame of Horty [7].

Definition 1 (Utilitarian frame). *A $utilitarian\ frame$ is a tuple $\langle W, A, Choice, \leq\rangle$, where W is a nonempty set of possible worlds, A is a finite set of agents, $Choice$ is a choice function, and $\leq$, represents the preference of the group A, is a reflexive and transitive relation on W.*

The choice function $Choice$ is a function from the power set of A to the power set of the power set of W, *i.e.* $Choice : \wp(A) \mapsto \wp(\wp(W))$. $Choice$ is built from the individual choice function $IndChoice$: $A \mapsto \wp(\wp(W))$. The $IndChoice$ must satisfy the following three conditions:

(1) for each $i \in A$ it holds that $IndChoice(i)$ is a partition of W;
(2) let $A = \{1, ..., n\}$, for every $x_1 \in IndChoice(1), \ldots, x_n \in IndChoice(n)$, $x_1 \cap \ldots \cap x_n \neq \emptyset$;

We call a function s: $A \mapsto \wp(W)$ a selection function if for each $i \in A$, $s(i) \in IndChoice(i)$. Let $Selection$ be the set of all selection functions, for any $G \subseteq A$, if $G \neq \emptyset$, we define $Choice(G) = \{\bigcap_{i\in G} s(i) : s \in Selection\}$. If $G = \emptyset$, we define $Choice(G) = \{W\}$. Here our utilitarian frame simplifies STIT frame in the sense that we restrict STIT frame to a single moment, therefore the concept of history in STIT frame is omitted.

Having defined utilitarian frames, we are ready to define *preferences over sets of possible worlds*.

Definition 2 (preferences over sets of worlds via strong lifting [14]). *Let $X, Y \subseteq W$ be two sets of worlds from a utilitarian frame. Then $X \preceq^s Y$(Y is $weakly\ preferred$ to X) if and only if*

(1) for each $w \in X$, for each $w' \in Y$, $w \leq w'$ and

(2) there exists some $v \in X$, some $v' \in Y$, $v \leq v'$.

$X \prec^s Y$ (Y is $strongly\ preferred$ to X) if and only if $X \preceq^s Y$ and it is not the case that $Y \preceq^s X$.

Definition 3 (dominance relation [7]). *Let F be a utilitarian frame. Let $G \subseteq A$ and $K, K' \in Choice(G)$. Then*

$K \preceq^s_G K'$ iff for all $S \in Choice(A - G)$, $K \cap S \preceq^s K' \cap S$.

$K \preceq^s_G K'$ is read as "K' weakly dominates K". From a decision theoretical perspective, $K \preceq^s_G K'$ means that no matter how other agents act, the outcome of choosing K' is no worse than that of choosing K. We use $K \prec^s_G K'$ as an abbreviation of $K \preceq^s_G K'$ but $K' \preceq^s_G K$ does not hold. If $K \prec^s_G K'$, we then say K' strongly dominate K.

Definition 4 (restricted choice sets [7]). *Let G be groups of agents from a utilitarian frame and X a set of worlds in the frame. Then*

$$Choice(G/X) = \{K : K \in Choice(G) \text{ and } K \cap X \neq \emptyset\}$$

Intuitively, $Choice(G/X)$ is the collection of those choices of group G that are consistent with condition X.

We now define a conditional dominance relation over agent's choice. The intuition is: to compare whether the agent's choice K is dominated by K' under the condition X, we only need to consider other agents' choices which are consistent with the condition X and at least one of K and K'.

Definition 5 (conditional dominance [14]). *Let G be groups of agents from a utilitarian frame and X a set of worlds in the frame. Let K, $K' \in Choice(G/X)$. Then*

$$K \preceq^s_{G/X} K' \text{ iff for all } S \in Choice((A-G)/(X \cap (K \cup K'))),\ K \cap X \cap S \preceq^s K' \cap X \cap S.$$

$K \preceq^s_{G/X} K'$ is read as "K' weakly dominates K under the condition of X". And we will use $K \prec^s_{G/X} K'$, read as "K' strongly dominates K under the condition of X", to express $K \preceq^s_{G/X} K'$ and it is not the case that $K' \preceq^s_{G/X} K$.

Definition 6 (Optimal and conditional optimal [7]). *Let G be a group of agents from a utilitarian frame,*

- $Optimal^s_G$ = $\{K \in Choice(G)$: *there is no* $K' \in Choice(G)$ *such that* $K \prec^s_G K'\}$.
- $Optimal^s_{G/X}$ = $\{K \in Choice(G/X)$: *there's no* $K' \in Choice(G/X)$ *such that* $K \prec^s_{G/X} K'\}$.

As in traditional modal logic, a model is a frame plus a valuation.

Definition 7 (utilitarian model). *A utilitarian model M is an ordered pair $\langle F, V\rangle$ where F is a utilitarian frame and V a valuation that assigns to each atomic proposition $p \in P$ a set of worlds $V(p) \subseteq W$.*

In the semantic of UDL, the optimal choices and conditional optimal choices are used to interpret the deontic operators.

Definition 8 (truth conditions). *Let $M =< \boldsymbol{F}, V >$ be a utilitarian model. Let $w \in W$ and let $\varphi, \psi \in L$. Then*

(1) $M, w \models p$ *iff $w \in$V(p);*
(2) $M, w \models \neg\varphi$ *iff it is not the case that* $M, w \models \varphi$*;*
(3) $M, w \models \varphi \wedge \psi$ *iff* $M, w \models \varphi$ *and* $M, w \models \psi$*;*
(4) $M, w \models \bigcirc_G \varphi$ *iff* $K \subseteq ||\varphi||$ *for each* $K \in Optimal^s_G$*;*
(5) $M, w \models \bigcirc_G(\varphi/\psi)$ *iff* $K \subseteq ||\varphi||$ *for each* $K \in Optimal^s_{G/\psi}$.

Here $\|\varphi\| = \{w \in W : M, w \models \varphi\}$.

We say φ is true in the world w of a utilitarian model M if $M, w \models \varphi$. Just like in the standard modal logic (for instance, Blackburn [1]), we introduce the concept of validity as follows: a formula φ is valid ($\models \varphi$) if it is true at every world of every utilitarian model.

4 Pessimistic utilitarian deontic logic

For pessimistic utilitarian deontic logic, instead of strong lifting, we using pessimistic lifting.

Definition 9 (preferences over sets of worlds via pessimistic lifting). *Let $X, Y \subseteq W$ be two sets of worlds from a utilitarian frame. Then $X \preceq^p Y$ if and only if there exists $w \in X$, such that for all $w' \in Y$, $w \leq w'$. $X \prec^p Y$ if and only if $X \preceq^p Y$ and it is not the case that $Y \preceq^p X$.*

Proposition 1. *Let X and Y be sets of worlds from a utilitarian frame. Then:*

1. *If $X \preceq^p Y$ and $Y \preceq^p Z$, then $X \preceq^p Z$.*
2. *If $X \preceq^p Y$ and $Y \prec^p Z$, then $X \prec^p Z$.*
3. *If $X \prec^p Y$ and $Y \preceq^p Z$, then $X \prec^p Z$.*

Proof. Here we prove the first two items, the third case is similar.

1. Assume $X \preceq^p Y$ and $Y \preceq^p Z$, then ther exist $u \in X$ such that for all $v \in Y$, $u \leq v$. There exist $v' \in Y$ such that for all $w \in Z$, $v' \leq w$. Then we have $u \leq v'$. For an arbitrary $w' \in Z$, we have $v' \leq w'$. Therefore $u \leq w'$.
2. Assume $X \preceq^p Y$ and $Y \prec^p Z$. From item *1.* we know $X \preceq^p Z$. To prove it is not the case that $Z \preceq^p X$, we assume otherwise. Then from $Z \preceq^p X$ and $X \preceq^p Y$ we deduce $Z \preceq^p Y$, contradicts to $Y \prec^p Z$. □

Proposition 1 states that the relation of preference over sets of worlds via pessimistic lifting is transitive. It is worth knowing that this proposition is crucial. Only with transitivity, we can properly define the concept of dominance and optimality. Another point we need to pay attention here is although we use the symble $\preceq^p$, normally we don't have $X \preceq^p X$ when X is not a singleton.

The definition of dominance ($\preceq^p_G$), conditional dominance ($\preceq^p_{G/X}$), optimal ($Optimal^p_G$) and conditional optimal ($Optimal^p_{G/X}$) in the pessimistic setting are obtained by simply changing $\leq^s$ to $\leq^p$ of their strong version counterpart.

Proposition 2. *Let G be groups of agents from a utilitarian frame, and let K, K', $K'' \in Choice(G)$. Then:*

1. *If $K \preceq^p_G K'$ and $K' \preceq^p_G K''$, then $K \preceq^p_G K''$.*
2. *If $K \preceq^p_G K'$ and $K' \prec^p_G K''$, then $K \prec^p_G K''$.*
3. *If $K \prec^p_G K'$ and $K' \preceq^p_G K''$, then $K \prec^p_G K''$.*

Proof. Here we prove the first two cases. The third case is similar.

1. Assume $K \preceq^p_G K'$ and $K' \preceq^p_G K''$, then for all $S \in Choice(A - G)$, $K \cap S \preceq^p K' \cap S$, $K' \cap S \preceq^p K'' \cap S$. Then by Proposition 1 we have $K \cap S \preceq^p K'' \cap S$. Hence $K \preceq^p_G K''$.
2. Similar to the proof of item 2 of Proposition 1. □

Proposition 3. *Let G be a group of agents from a utilitarian frame and X a set of worlds in the frame. Let K, K', $K'' \in Choice(G/X)$. Then the following holds,*

1. *If* $K \preceq_{G/X} K'$ *and* $K' \preceq_{G/X} K''$*, then* $K \preceq_{G/X} K''$.
2. *If* $K \preceq_{G/X} K'$ *and* $K' \prec_{G/X} K''$*, then* $K \prec_{G/X} K''$.
3. *If* $K \prec_{G/X} K'$ *and* $K' \preceq_{G/X} K''$*, then* $K \prec_{G/X} K''$.

Proof. See Proposition 12 of Sun [14].

Proposition 4. $O(A/B) \wedge O(A/C) \rightarrow O(A/(B \vee C))$ *is not valid in pessimistic utilitarian deontic logic.*

Proof. It's sufficient to construct a counter-example. Let $M = (W, Agent, Choice, \leq, V), W = \{w_1, \ldots, w_6\}, Agent = \{1, 2\}, Choice(\{1\}) = \{\{w_1, w_2\}, \{w_3, w_4\}, \{w_5, w_6\}\}$, $Choice(\{2\}) = \{W\}$. $w_3 \leq w_4 \leq w_1 \leq w_2 \leq w_5 \leq w_6$. $V(p) = \{w_1, w_2\}$, $V(q) = \{w_1, w_3, w_5\}$. It can be verified that M is a utilitarian model. Moreover, we have $M, w \models \bigcirc(p/q) \wedge \bigcirc(p/\neg q)$ but $M, w \not\models \bigcirc(p/(q \vee \neg q))$.

4.1 Application: an analysis of the miners puzzle

The miners puzzle can be represented by the following figure:

block_neither	w_1 *in_A* (9)	w_2 (9) *in_B*
block_B	w_3 *in_A* (0)	w_4 (10) *in_B*
block_A	w_5 *in_A* (10)	w_6 (0) *in_B*

Figure 4.1: $W = \{w_1, \ldots, w_6\}$, $w_3 \approx w_6 \leq w_1 \approx w_2 \leq w_4 \approx w_5$.

We assume nature and rescuer are the only two agents involved in this scenario. We further assume nature has only one dummy choice the outcome of which is the set of all possible worlds. The rescuer has three choices: *block_neither*, *block*_A and *block*_B. According to the pessimistic semantics, *block_neither* is the only dominated choice. For each choice there are two outcomes. Therefore in total there are six possible worlds. The preference of worlds are determined by the number of miners saved in that world. Then according to the pessimistic semantics, *block_neither* is the only optimal choice. So we can draw the prediction that "the rescuer ought to block neither". Moreover, given the condition of miners being in A, *block*_A becomes the only conditional optimal choice. Hence we have "if the miners are in A, then the rescuer ought to block A". The case for miners being in B are similar. Although we have both "if the miners are in A, then the rescuer ought to block A" and "if the miners are in B, then the rescuer ought to block B", by Proposition 4 we can avoid the prediction that "the rescuer ought to block either A or B". Therefore the pessimistic utilitarian deontic logic gives a viable solution to the miners puzzle.

Compared to those approaches reviewed in Section 2, our STIT based approach has stronger expressive power. We have agents and action modality in our language. This gives considerable expressive power already. Just like Horty [7] and Broersen [2], our framework can easily be extended to involve temporal modality, which further increase the expressive power.

5 Conclusion and future work

This paper develops a new STIT based deontic logic, pessimistic utilitarian deontic logic, capable of analyzing the miners puzzle. The key idea of the semantics of this logic is: one set of possible worlds is better another set of possible worlds iff the worst world in the first set is better than the worst world in the second. This semantics gives write predictions in the miners scenario meanwhile blocks the problematic prediction.

Concerning future works, an axiomatisation of pessimistic utilitarian deontic logic is worthy investigating. A second potential extension is to use other STIT operators, for example the deliberative STIT [8] and X-STIT [2].

References

1. Patric Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge University Press, Cambridge, 2001.
2. Jan Broersen. Deontic epistemic *stit* logic distinguishing modes of mens rea. *Journal of Applied Logic*, 9(2):127 – 152, 2011.
3. Fabrizio Cariani, Magdalena Kaufmann, and Stefan Kaufmann. Deliberative modality under epistemic uncertainty. *Linguistics and Philosophy*, 36(3):225–259, 2013.
4. Jennifer Carr. Deontic modals without decision theory. *Proceedings of Sinn und Bedeutung*, 17:167–182, 2012.
5. Nate Charlow. What we know and what to do. *Synthese*, 190(12):2291–2323, 2013.
6. Dov Gabbay, Livio Robaldo, Xin Sun, Leendert van der Torre, and Zohreh Baniasadi. A new analysis of the miners puzzle: A beth semantics approach. In *DEON2014*.
7. John Horty. *Agency and Deontic Logic*. Oxford University Press, New York, 2001.
8. John Horty and Nuel Belnap. The deliberative stit: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24:583–644, 1995.
9. Niko Kolodny and John MacFarlane. Iffs and oughts. *Journal of Philosophy*, 107(3):115–143, 2010.
10. Barteld Kooi and Allard Tamminga. Conditioanl obligations in strategic situations. In M. Singh G. Boella, G Pigozzi and H. Verhagen, editors, *3rd International Workshop on Normative Multiagent Systems(NorMAS 2008)*, pages 188–200, Luxembourg, July 2008.
11. Barteld Kooi and Allard Tamminga. Moral conflicts between groups of agents. *Journal of Philosophical Logic*, 37:1–21, 2008.
12. Angelike Kratzer. The notional category of modality. In H. J. Eikmeyer and H. Rieser, editors, *Words, worlds, and Contexts: New Approaches in World Semantics*. Berlin: de Gruyter, 1981.
13. Jérôme Lang and Leendert van der Torre. From belief change to preference change. In *Proceedings of Eighteenth European Conference on Artificial Intelligence (ECAI2008)*, pages 351–355, 2008.
14. Xin Sun. Conditional ought, a game theoretical perspective. In J. Lang H. van Ditmarsch and S. Ju, editors, *Logic, Rationality, and Interaction: Proceedings of the Thire International Workshop*, pages 356–369, Guangzhou, China, 2011.
15. Allard Tamminga. Deontic logic for strategic games. *Erkenntnis*, pages 1–18, 2011 Nov. 24.
16. Malte Willer. A remark on iffy oughts. *Journal of Philosophy*, 109(7):449461, 2012.